

# Statistical validation of high-dimensional models of growing networks

Matúš Medo

*Physics Department, Chemin du Musée 3, University of Fribourg, 1700 Fribourg, Switzerland*

(Received 7 November 2013; published 3 March 2014)

The abundance of models of complex networks and the current insufficient validation standards make it difficult to judge which models are strongly supported by data and which are not. We focus here on likelihood maximization methods for models of growing networks with many parameters and compare their performance on artificial and real datasets. While high dimensionality of the parameter space harms the performance of direct likelihood maximization on artificial data, this can be improved by introducing a suitable penalization term. Likelihood maximization on real data shows that the presented approach is able to discriminate among available network models. To make large-scale datasets accessible to this kind of analysis, we propose a subset sampling technique and show that it yields substantial model evidence in a fraction of time necessary for the analysis of the complete data.

DOI: [10.1103/PhysRevE.89.032801](https://doi.org/10.1103/PhysRevE.89.032801)

PACS number(s): 89.75.Hc, 64.60.aq, 87.23.Ge, 02.50.—r

## I. INTRODUCTION

Complex networks are used to represent and analyze a wide range of systems [1–3]. Models of complex networks usually aim for simplicity and attempt to keep the number of parameters as low as possible. However, real data is more complex than any simple model which makes it difficult to draw clear links between data and models. To capture the increasingly available massive real data [4], we need high-dimensional models where the number of parameters grows with the number of nodes. An example of such a model is the latent space model [5] where nodes are assigned independent and identically distributed vectors and the probability of a link connecting two nodes depends only on the distance of their vectors.

While there are plenty of simple (and not so simple) network models, little is known as to which of them are really supported by data. While calibration of complex network models often uses standard statistical techniques, their validation is typically based on comparing their aggregate features (such as the degree distribution or clustering coefficient—see Refs. [6,7] for detailed accounts on network measurements) with what is seen in real networks (see Refs. [8,9] for recent examples of this approach). The focus on aggregate quantities naturally reduces the discriminative power of model validation which is often further harmed by the use of inappropriate statistical methods [10]. As a result, we still lack knowledge of what is to date the best model explaining the growth of the scientific citation network, for example.

We argue that network models need to be evaluated by robust statistical methods [11,12], especially by those that are suited to high-dimensional models [13]. This is exemplified in [14] where various low-dimensional microscopic mechanisms for evolution of social networks are compared on the basis of their likelihood of generating the observed data. Prohibitive computational complexity of maximum likelihood estimation is often quoted as a reason for its limited use in the study of real world complex networks [15]. However, as we shall see here, even small subsets of data allow to discriminate between models and point clearly to those that are actually supported by the data. This, together with the ever-increasing computational power at our disposal, opens the door to the likelihood analysis of complex network models.

We analyze here a recent network growth model [16,17] which naturally leans itself to high-dimensional analysis. This model generalizes the classical preferential attachment (PA; often referred to as the Barabási–Albert model in the complex networks literature) [17, Secs. 7, 8] by introducing node relevance which decays in time and co-determines (together with node degree) the rate at which nodes acquire new links. If either the initial relevance values or the functional form of the relevance decay are heterogeneous among the nodes, this model is able to produce various realistic degree distributions. By contrast to Ref. [18], which modifies preferential attachment by introducing an additive heterogeneous term, in Ref. [16] relevance combines with degree in a multiplicative way which means that once it reaches zero, the degree growth stops. This makes the model an apt candidate for modeling information networks where information items naturally lose their pertinence with time and the growth of their degree eventually stops (see Ref. [19] for a review of work on temporal networks). This model has been recently used to quantify and predict citation patterns of scientific papers [20].

Before methods for high-dimensional parameter estimation are applied to real data, we calibrate and evaluate them on artificial data where one has full control over global network parameters (size, average degree, etc.) and true node parameter values are known. For simplicity, we limit our attention to the case where the functional form of relevance decay is the same for all nodes and only the initial relevance values differ. We present here various estimation methods and evaluate their performance. Plain maximum likelihood [12, Chapter 7] produces unsatisfactory results, especially in the case of sparse networks which are commonly seen in practice. We enhance the method by introducing an additional term which suppresses undesired correlation between node age and estimates of initial relevance. We then introduce a mean-field approach which allows us to reduce high-dimensional estimation to a low-dimensional one. Calibration and evaluation of these parameter-estimation methods is done on artificial data. Real data is then used to employ the established framework and compare the statistical evidence for several low- and high-dimensional network models on the given data. Analysis of small subsets of input data is shown to

efficiently discriminative among the available models. Since this work focuses on model evaluation, estimated parameter values are thus of secondary importance to us. Necessary conditions for obtaining precise estimates and the potential risk of large errors [21] are therefore left for future research (see Sec. VI).

## II. MODEL

The original model of preferential attachment with relevance decay (PA-RD) has been formulated for an undirected network where the initial node degree is nonzero because of links created by the node on its arrival [16]. To allow zero-degree nodes to collect links, some additive attractiveness or random node selection need to be introduced. When these two mechanisms are combined with PA-RD, the probability that a new link created at time  $t$  attaches to node  $i$  can be written as

$$P(i, t) = \lambda \frac{R_i(t)[k_i(t) + A]}{\sum_{j=1}^{n(t)} R_j(t)[k_j(t) + A]} + \frac{1 - \lambda}{n(t)}. \quad (1)$$

Here  $k_i(t)$  and  $R_i(t)$  are degree and relevance of node  $i$  at time  $t$ , respectively,  $n(t)$  is the number of nodes present at time  $t$ , and  $A$  is the additive attractiveness term. Finally,  $\lambda$  is the probability that the node is chosen by the PA-RD mechanism; the node is chosen at random with the complementary probability  $1 - \lambda$ . When  $A = 0$  and  $\lambda = 1$ , a node of zero degree will never attract new links. Equation (1) can be used to model a monopartite network where nodes link to each other as well as a bipartite network where nodes link to each other as well as a bipartite network where one set of nodes is unimportant and we can thus speak of outside links attaching to nodes. For example, one can use the model to describe the dynamics of item popularity in a user-item bipartite network representing an e-commerce system [22,23].

There are now two points to make. First, the model is invariant with respect to the rescaling of all relevance values,  $R_i(t) \rightarrow \xi R_i(t)$ . This may lead to poor convergence of numerical optimization schemes because  $R_i(t)$  values can drift in accord without affecting the likelihood value. The convergence problems can be avoided by imposing an arbitrary normalization constraint on the relevance values as we do below. Second,  $A$  and  $\lambda$  act in the same direction: they introduce randomness in preferential attachment-driven network growth (in particular, as  $A \rightarrow \infty$  and/or  $\lambda \rightarrow 0$ , preferential attachment loses all influence). One can therefore expect that  $A$  and  $\lambda$  are difficult to be simultaneously inferred from the data. This is especially true for the original preferential attachment without decaying relevance. If node relevance decays to zero, node attraction due to  $A$  eventually vanishes while the random-attachment part proportional to  $\lambda$  remains—it is therefore possible, at least in principle, to distinguish between the two effects. To better focus on the high-dimensional likelihood maximization of node parameters, we assume  $\lambda = 1$  in all our simulations.

The PA-RD model has been solved in Ref. [16] for a case where  $\lambda = 1$ ,  $A = 0$ , and the initial degree of all nodes equal to one. It was further assumed that  $T_i := \int_0^\infty R_i(t)dt$  is finite for all nodes and the distribution of  $T$  values among the nodes,  $\varrho(T)$ , decays exponentially or faster. The probability normalization term  $\sum_j R_j(t)[k_j(t) + A]$  then eventually fluctuates

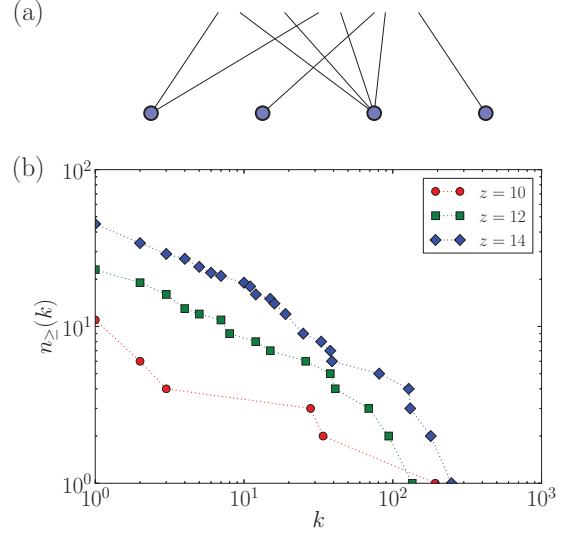


FIG. 1. (Color online) (a) Illustration of a bipartite network where only links' target nodes are of interest. (b) Sample degree distributions of networks produced according to Sec. II A.

around a stationary value  $\Omega^*$  and the expected final degree of node  $i$  can be written as  $\langle k_i^F \rangle = \exp(T_i/\Omega^*)$ . It has been shown that the network's degree distribution, shaped mainly by  $\varrho(T)$ , can take on various forms including exponential, log-normal, and power-law.

### Description of artificial data

We begin by describing bipartite network data with temporal information. We consider a simplified bipartite case where links arrive from outside and thus only their target nodes matter—see Fig. 1(a) for illustration. Links are numbered with  $l = 1, \dots, E$  and the times at which they are introduced are  $t_1 \leq t_2 \leq \dots \leq t_E$ . Nodes are numbered with  $i = 1, \dots, N$  and the times at which they are introduced are  $\tau_1 \leq \tau_2 \leq \dots \leq \tau_N$ . At time  $t$ , there are  $n(t)$  target nodes in the network. Degree of node  $i$  at time  $t_l$  when link  $l$  is added is  $k_i(t_l)$  and the target node of link  $l$  is  $n_l$ . The average node degree is  $z := E/N$  (the factor of two is missing here because we consider a bipartite network where  $E$  edges point to  $N$  nodes of interest).

We use the PA-RD model to create artificial networks with well-defined properties. There are initially  $n_I$  nodes with zero degree. After every  $\Delta T$  time steps, a new node of zero degree is introduced in the network. In each time step, one new link is created and chooses its target node according to Eq. (1). The network growth stops once there are  $E_F = \lceil zn_I/(1 - z/\Delta T) \rceil$  links and  $n_F = n_I + \lceil E_F/\Delta T \rceil$  nodes in the network. At that point, the average node degree is approximately  $z$ . It must hold that  $z < \Delta T$ ; in the opposite case, the average degree  $z$  cannot be achieved because new nodes dilute the network too fast. Each node has the relevance decay function  $R_i(t) = I_i \exp[-(t - \tau_i)/\Theta]$  where  $\Theta$  is the decay time scale and  $I_i$  is the initial relevance of node  $i$ . Initial relevance values are drawn from the exponential distribution  $f(I) = e^{-I}$ . When the decay parameter  $\Theta$  is sufficiently high, this setting produces broad degree distributions [16] which are similar to distributions often seen in real information networks [3,

Chapter 4]. We use  $n_l = 10$ ,  $\Delta T = 16$ ,  $\Theta = 50$ ,  $A = 1$ , and  $\lambda = 0$  for all artificial networks studied here; their sample degree distributions are shown in Fig. 1(b).

### III. PARAMETER ESTIMATION METHODS

#### A. Maximum likelihood estimation

We first use the standard maximum likelihood estimation (MLE) to estimate parameters of the PA-RD model [11]. A generic form of log-likelihood of realization  $\mathcal{D}$  for a network growth model  $\mathcal{M}$  has the form

$$\ln \mathcal{L}(\mathcal{D}|\mathcal{M}) = \sum_{l=1}^E \ln P(n_l, t_l|\mathcal{M}), \quad (2)$$

where  $P(n_l, t_l|\mathcal{M})$  is the probability of link  $l$  arriving at node  $n_l$  at time  $t_l$  under model  $\mathcal{M}$ . It is convenient to transform this quantity into log-likelihood per link by dividing it with the number of links,  $\ln \mathcal{L}(\mathcal{D}|\mathcal{M})/E$ . For model  $\mathcal{M}$  represented by its attachment probability  $P(n_l, t_l|\mathcal{M})$  and a vector of model parameters  $\mathbf{p}$ , log-likelihood can be maximized with respect to these parameters and yields their estimates  $\tilde{\mathbf{p}}$ .

Given a network realization obtained with Eq. (1), there are several parameters to estimate: initial relevance values of all nodes, additional attractiveness term  $A$ , and parameters of the relevance decay function. (Note that we make the estimation task easier by assuming that the functional form of relevance decay is known.) Greedy (uphill) maximization of log-likelihood is made possible by the profile of the likelihood function which does not feature multiple local maxima in the space of initial relevance values (see Sec. A for an explanation). Starting from a random initial guess, we sequentially update all model parameters by quadratic extrapolation and repeat this process until the difference between new and old estimates is less than some sufficiently small threshold (we use  $10^{-3}$  here). Due to the scale-invariance of relevance values, they can be normalized after each iteration so that their average is one, which improves convergence. While each evaluation of log-likelihood is time consuming and this straightforward approach is thus computationally expensive, it is often, as we shall show, viable.

#### B. Mean-field approximation to maximum likelihood estimation

As mentioned in Sec. II, when the number of nodes is large and their relevance decays to zero, fluctuations of the denominator in Eq. (1) become small and one can therefore replace it with a constant term  $\Omega^*$ . This mean-field approximation decouples the dynamics of nodes which then compete for new links with the external field  $\Omega^*$  instead of competing with the other nodes present in the system. Equation (1) then simplifies to

$$P(i, t) = \frac{R_i(t|\boldsymbol{\eta}_i)[k_i(t) + A]}{\Omega^*}, \quad (3)$$

where  $\boldsymbol{\eta}$  is a vector of parameters of node  $i$  and we again assume  $\lambda = 1$ . In our case, the initial relevance value  $I_i$  is the only node-specific parameter and thus  $\boldsymbol{\eta}_i = (I_i)$ . Since  $\Omega^*$  is the same for all nodes, we can subsume it in  $I_i$  due to the aforementioned scale invariance. The likelihood function for node  $i$  is then constructed by evaluating all links created after

this node has been introduced in the network. For link  $l$ , we assess whether the link points to node  $i$  (then  $\delta_{i, n_l} = 1$ ) or not (then  $\delta_{i, n_l} = 0$ ). We get

$$\ln \mathcal{L}_i(\mathcal{D}|\boldsymbol{\eta}_i) = \sum_{\substack{l=1 \\ t_l \geq t_i}}^E \ln \{P(i, t_l)\delta_{i, n_l} + [1 - P(i, t_l)](1 - \delta_{i, n_l})\}, \quad (4)$$

where we ignore links that are older than node  $i$ . This function can be maximized with respect to  $\boldsymbol{\eta}_i$  for any given  $A$ . Global model parameters such as, in our case,  $A$  and the time scale of relevance decay  $\Theta$  can be estimated by minimizing  $\sum_i \ln \mathcal{L}_i(\mathcal{D}|\tilde{\boldsymbol{\eta}}_i)$  with respect to them (estimates  $\tilde{\boldsymbol{\eta}}_i$  then need to be updated to reflect new values of the global parameters).

The mean-field approximation to the maximum likelihood estimation (MF-MLE) makes it easy to change the functional form of relevance decay  $R(t)$  for any individual node and thus classify their behavior (see Ref. [23, Chapter 8] for more information on classification problems). While we do not pursue this direction here, it is of particular significance to the analysis of real data where various behavioral classes of nodes are likely to coexist. Also, the vector of node parameters can be easily extended by, for example, making the decay time  $\Theta$  node dependent, while still maintaining the low-dimensional nature of the resulting likelihood optimization.

### IV. ESTIMATION EVALUATION

To evaluate various estimation methods, we assess the maximal likelihood that they are able to achieve. Parameter estimation is simplified by assuming that the functional form of relevance decay is known and only model parameters  $A, \Theta, \{I_i\}_i$  are to be estimated. Since the true parameter values are available to us, we also measure Pearson's correlation between true values  $I$  and their estimates  $\tilde{I}$ ,  $r(I, \tilde{I})$  (the higher the value, the better the estimates). In evaluating this correlation, nodes with final degree four and less are excluded because their estimates are too noisy due to the lack of data. The advantage of using Pearson's correlation to measure the accuracy of estimates lies in its invariance with respect to rescaling of  $\tilde{I}_i$  which fits well with the scale-invariance of the PA-RD model itself. The accuracy of estimates of  $A$  and  $\Theta$  is measured as well.

Simulations reveal that MLE sometimes converges to estimates which are far from the true parameter values. To explain the reason for this behavior, Fig. 2 shows the results of constrained likelihood maximization where we artificially fix  $A$  and  $\Theta$  at various values, many of which are far from the true values  $A = 1$  and  $\Theta = 50$ . The corresponding maximal log-likelihood values exhibit a shallow maximum in  $A$  with the optimal value 2.7 lying significantly above the true value 1. Worse, the maximum in  $\Theta$  is nonexistent: as  $\Theta$  increases, log-likelihood increases too and saturates at a value which is maintained also in the limit  $\Theta \rightarrow \infty$  (i.e., no relevance decay). Resulting  $\tilde{\Theta}$  thus depends on the initial values of model parameters and the procedure in which they are iteratively improved in the search for maximal likelihood. While Fig. 2 shows results for one network realization, the same behavior can be seen for all realizations of the input artificial network. Inspection of the initial relevance values estimated for large

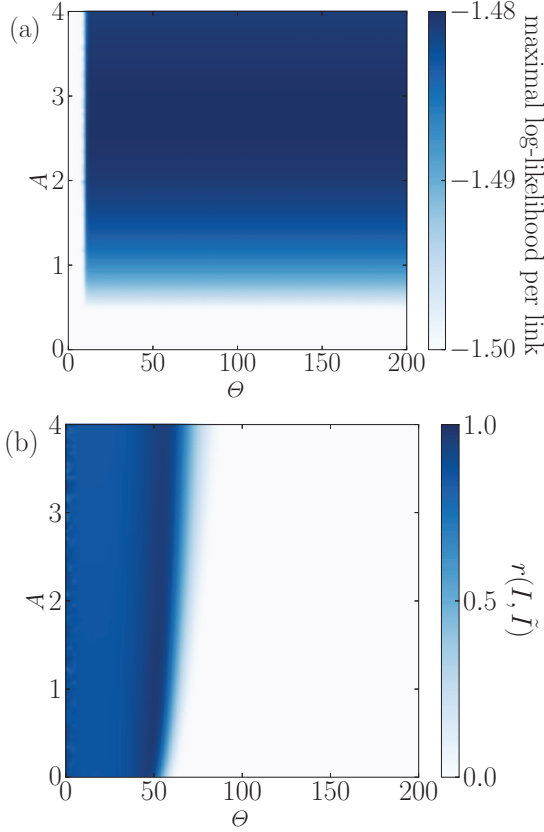


FIG. 2. (Color online) Results of a constrained MLE procedure where  $A$  and  $\Theta$  are fixed at various values and log-likelihood is thus maximized only with respect to the initial relevance of each node. These results were obtained for one realization of the artificial network model with  $z = 12$  which corresponds to  $n_F = 30$  and  $E_F = 480$ .

$\Theta$  makes it clear that the lack of relevance decay is then compensated by later nodes being assigned higher initial relevance than earlier ones. As a result, MLE estimates then do not reflect the true initial relevance values but rather the order in which nodes are introduced in the network. This is demonstrated by the second panel of Fig. 2 where  $r(I, \tilde{I})$  reaches maximum for  $\Theta$  close to the true value of 50 and then quickly drops to negative values for larger values of  $\Theta$ . The negative correlation values are observed here because in this particular network realization, node arrival times are negatively correlated with their initial relevance values. The overall maximum of  $r(I, \tilde{I})$  lies at  $A = 0.94$  and  $\Theta = 51$ .

The problem of excessive estimated decay time  $\Theta$  can be solved by introducing an additional term in log-likelihood with the aim to penalize solutions with high  $\Theta$ . This is similar to regularization schemes such as the least absolute shrinkage and selection operator (LASSO) [24] which are often used to constraint solutions in high-dimensional optimization problems [13]. We choose here to maximize

$$\frac{1}{E} \ln \mathcal{L}_i(\mathcal{D}|I_i, A, \Theta) - \omega r(\tau, \tilde{I})g[r(\tau, \tilde{I})], \quad (5)$$

where  $g(x) = x$  for  $x > 0$  and 0 otherwise; the additional term penalizes positive correlation between nodes' arrival

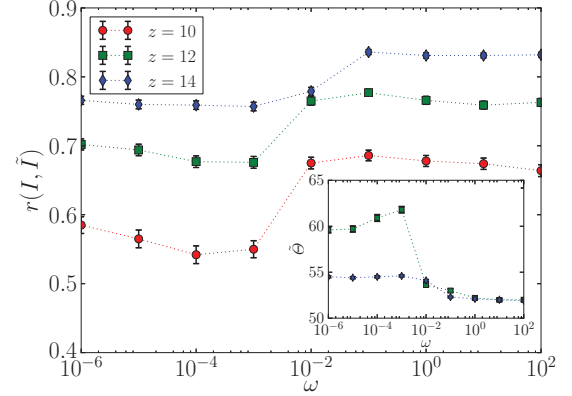


FIG. 3. (Color online) Impact of the log-likelihood penalization term given in Eq. (5) as a function of  $\omega$  for various values of  $z$ . Mean values and their standard errors were obtained by estimation in 1000 independent network realizations (the same applies to Table I and Fig. 4). Estimates of  $\Theta$  are large for  $z = 10$  and thus missing in the inset.

times and their estimated initial relevances. As shown in Fig. 3, MLE estimates with the correlation term  $r(\tau, \tilde{I})$  are superior to the original ones over a broad range of  $\omega$ . The difference is particularly large for sparse networks where standard MLE strongly overestimates  $\Theta$ . Note that unlike for large parts of Fig. 2(b), the average correlation value  $r(I, \tilde{I})$  in Fig. 3 is positive even when  $\Theta$  is large. This is because while Fig. 2 presents outcome for a single network realization, Fig. 3 averages over many of them and the resulting average correlation is positive.

We proceed now to a direct comparison of the discussed estimation methods. As can be seen in Table I, all methods produce log-likelihood of a comparable magnitude but their parameter estimates (only  $\tilde{A}$  and  $\tilde{\Theta}$  are shown here) differ. Notably, while  $\tilde{\Theta}$  is close to the true value, the error of  $\tilde{A}$  is substantial. One can say that MLE tends to overestimate the tendency to random connections (as  $A$  grows, the influence of preferential attachment vanishes) in this artificial system. We found no signs of this error vanishing with the data size which can be probably attributed to the network growth which constantly injects new nodes with zero degree in the system. The highest correlation between  $I$  and  $\tilde{I}$  is achieved with penalized MLE estimation when  $\omega = 1$ . Performance of the methods is further illustrated in Fig. 4 as a function of  $z$ . As expected,  $r(I, \tilde{I})$  increases with  $z$  for both exact MLE methods and penalized estimations always outperform the unpenalized ones. The behavior is different for results obtained with the mean-field MLE whose quality slowly deteriorates as

TABLE I. Estimates obtained with respective methods for  $z = 14$  (which corresponds to  $n_F = 80$ ,  $E_F = 1120$ ). Numbers in brackets report uncertainty of the last digit given by the standard error of the mean.

Method	$\ln \mathcal{L}/E$	$\tilde{A}$	$\tilde{\Theta}$	$r(I, \tilde{I})$
MLE, $\omega = 0$	-1.350(5)	2.86(2)	54.4(1)	0.767(5)
MLE, $\omega = 1$	-1.350(5)	2.85(2)	51.8(1)	0.835(3)
MF-MLE	-1.381(5)	6.09(9)	65.2(2)	0.722(2)



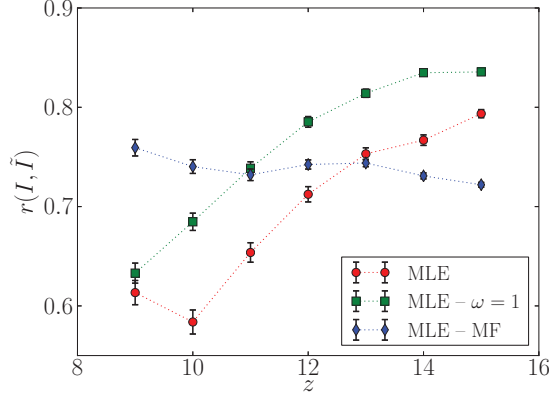


FIG. 4. (Color online) Performance of estimation methods vs mean degree  $z$  of artificial networks.

$z$  increases. Obviously, further improvements are necessary to make this otherwise promising method applicable in practice.

## V. ANALYSIS OF REAL DATA

To illustrate the potential of high-dimensional statistical analysis of network data, we finally apply it to real data and compare the level of support which it gives to various network models. The analyzed data originates from the interdisciplinary physics community web site *Econophysics Forum* (see Ref. [25]) which is run by the research group of Yi-Cheng Zhang at the University of Fribourg since 1998. We parsed server web log files collected from 6th July 2010 until 31st March 2013 (a time span of 1000 days). Activities of web bots and other automated access were removed from the data. While web logs contain all user actions on the web site, we kept only entries corresponding to downloads of papers posted on the *Econophysics Forum*. The corresponding user-paper bipartite network consists of 844 paper-nodes and their 24 581 links [26]. As expected, the degree distribution of paper-nodes is broad (the maximal degree of 741 is much greater than the average degree of 29), making this data a good candidate for being explained by preferential attachment or related models.

We use this data to evaluate two low-dimensional and four high-dimensional models. The low-dimensional models are random attachment to an existing node (RAND) and the standard preferential attachment (PA). The high-dimensional models are preferential attachment with heterogeneous (node-dependent) additive term (PA-H), preferential attachment with heterogeneous and decaying additive term (PA-HD) which

has been introduced in Ref. [18], preferential attachment with constant relevance (PA-R; such constant relevance is usually referred as fitness in past works [27]), and finally preferential attachment with relevance decay (PA-RD). The functional form of the probability of a new link attaching to an existing node at time  $t$ ,  $P_i(t)$ , is shown in Table II for each model. The form of  $A_i(t)$  suggested for PA-HD in Ref. [18] is generalized to  $A_i(t) = I_i(t) \exp\{-(t - \tau_i)/\Theta\}^\beta\}$  which in our case performs better than the original form without  $\beta$ . Note that Ref. [28] reports a similar behavior in the popularity growth of stories in *digg.com*. For simplicity, we assume a similar form of  $R_i(t)$  in PA-RD,  $R_i(t) = I_i(t) \exp\{-(t - \tau_i)/\Theta\}^\beta + R_\infty$ , which in fact roughly corresponds to the empirical relevance decay results presented in Ref. [16]. A nonvanishing absolute term  $R_\infty$  is needed here to allow for links occasionally attaching to old nodes. The log-normal decay form reported in Refs. [20,29] does not yield better fit in our case, perhaps as a result of immediate response of the *Econophysics Forum* users which makes the increasing relevance phase provided by log-normal curves unfitting. For PA-RD, we report results obtained with the penalization term ( $\omega = 1$ ) which, however, differ little from the results obtained with  $\omega = 0$ .

To maximize the likelihood functions we use the iterative extrapolating approach described in Sec. III A. This procedure is run ten times with independent random initial configurations; the best result obtained with each method is reported in Table II. In addition, the table shows also the number of model parameters  $k_M$  and the corrected Akaike information criterion

$$C_{AIC}(M) = -2 \ln[\max \mathcal{L}(\mathcal{D}|M)] + \frac{2k_M E}{E - k_M - 1}, \quad (6)$$

where the maximum is taken over the whole parameter space of model  $M$ .  $C_{AIC}(M)$  measures how well model  $M$  fits the data and corrects for a finite sample size [30]. It can be used to construct model weights [31] in the form

$$w_M \sim \exp \left\{ \left[ \min_{M'} C_{AIC}(M') - C_{AIC}(M) \right] / 2 \right\}, \quad (7)$$

where the proportionality factor is obtained by requiring the sum of all model weights to equal one. Finally, we report the values of global model parameters that maximize data likelihood for each model.

Our comparison of models contains several notable outcomes. First, both low-dimensional models are clearly insufficient to explain the data. In fact, preferential attachment yields only marginally better fit than random attachment. Second, high-dimensional models without time decay perform significantly worse than their counterparts with time decay.

TABLE II. Maximum likelihood estimates and model selection results for the *Econophysics Forum* data. Model weights  $w_M$  defined by Eq. (7) show that there is overwhelming evidence in favor of the model with decaying relevance (PA-RD).

Model $M$	$P_i(t)$	$k_M$	$\ln \mathcal{L}/E$	$C_{AIC}(M)$	$w_M$	$\bar{A}$	$\bar{R}_\infty$	$\bar{\Theta}$	$\bar{\beta}$
RAND	1	0	-5.805	285 364	0				
PA	$k_i(t) + A$	1	-5.767	283 519	0	137			
PA-H	$k_i(t) + A_i$	$N$	-4.641	229 931	0				
PA-HD	$k_i(t) + A_i(t)$	$N + 2$	-4.111	203 872	0			1.9	0.42
PA-R	$[k_i(t) + A]R_i$	$N + 1$	-4.641	229 905	0	1108			
PA-RD	$[k_i(t) + A]R_i(t)$	$N + 4$	-4.043	200 536	1	60	0.0088	3.2	0.64

This is not surprising because we fit the models to an information network where, as argued in Ref. [16], aging of nodes is of prime importance. Third, while the log-likelihood values obtained with PA-HD and PA-RD are both substantially better than those obtained for other models, the difference between them is big enough for the Akaike information criterion to assign an overwhelming weight to PA-RD. (The resulting weight of PA-HD, which has been truncated to zero in Table II, is around  $10^{-724}$ .)

For PA-RD, the effective lifetime corresponding to the obtained relevance decay parameters is

$$\langle t \rangle := \frac{\int_0^\infty t R(t) dt}{\int_0^\infty R(t) dt} = \frac{\tilde{\Theta} \Gamma(2/\tilde{\beta})}{\tilde{\beta} \Gamma(1 + 1/\tilde{\beta})} \approx 8 \text{ days}, \quad (8)$$

where we neglect  $R_\infty$  which is small, yet it formally causes the above-written expression to diverge. This lifetime well agrees with the fact that papers typically spend one week on the front page of the *Econophysics Forum*. The value of the additive term  $\tilde{A} \approx 60$  is relatively high in comparison with the average node degree of 29 which suggests that in the studied dataset, the influence of preferential attachment (i.e., attachment probability proportional to node degree) is relatively weak. An alternative explanation is that our assumed relevance decay function  $R(t)$  disagrees with the data and thus an increased proportion of “random” connections is necessary to model the data. A more detailed analysis is necessary to establish what is the real reason behind this apparent randomness.

Since likelihood computation is costly and during its maximization in numerous variables it needs to be carried out many times, obtaining the results presented in Table II on a standard desktop computer takes several hours. It is thus natural to ask whether significant evidence in favor of one of the models cannot be obtained by analyzing subsets of the data which would save considerable computational time. To this end, we evaluated weights of three representative models (PA, PA-HD, and PA-RD) on data subsets corresponding to time spans (which we refer to as subset lengths,  $L$ ) ranging from 4 to 100 days. We generated 1000 subsets for each  $L$  by choosing their starting day at random. Results shown in Fig. 5 demonstrate that while particularly short subsets

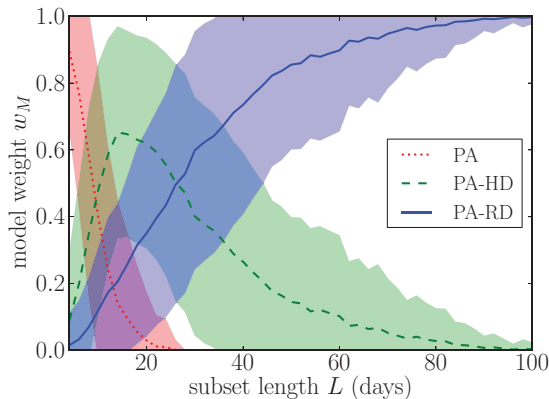


FIG. 5. (Color online) Model weights as a function of subset length for PA, PA-HD, and PA-RD. Lines show the mean weights and shaded areas mark their standard deviation based on 1000 random subsets drawn for each subset length.

favor the low-dimensional PA model, the situation quickly changes and this model is virtually eliminated as soon as  $L \gtrsim 30$ . Two high-dimensional models, which enjoy comparable support until  $L \lesssim 30$ , are clearly distinguished at  $L = 60$  and above. Meanwhile, evaluation of multiple small-scale subsets is fast: the computational time required for one likelihood maximization of PA-RD drops from 10 minutes for the whole 1000-day data spanning to 2 seconds for a 100-day subset. We can conclude that this approach allows us to efficiently discriminate between models even when no particularly efficient approach to likelihood maximization is available.

## VI. DISCUSSION

We studied the use of maximum likelihood estimation in analysis of high-dimensional models of growing networks. Artificially created networks with preferential attachment and decaying relevance [16] were used to show that a near-flat likelihood landscape makes the standard likelihood maximization rather unreliable and sensitive to the initial choice of model parameters. Introducing a penalization term effectively modifies the landscape and helps to avoid “wrong” solutions. The resulting MLE-based scheme outperforms the standard likelihood maximization for a wide range of model networks. On the other hand, both original and modified MLE overestimate the additive parameter  $A$  which is crucial in the early stage of a node’s degree growth. How to improve on that remains an open question.

We then tested the previously developed methods on real data where both preferential attachment and relevance decay are expected to play a role. In this part, the focus is on comparing various competing network models that may be used to explain the data. We show that the data shows overwhelming evidence in favor of one of the models and that sufficiently strong evidence can be achieved by studying small subsets of the data. Model evaluation by such subset sampling is of particular importance to large-scale datasets where straightforward likelihood maximization is prohibitively time consuming.

Up to now, models of complex networks have been appraised mostly by comparing aggregate characteristics of the produced networks (degree distribution or clustering coefficient, for example) with features seen in real data. The caveat of this approach is that many network characteristics are computed on static network snapshots and are thus of little use for the measurement of growing networks. Empirical node relevance [16] is designed especially for growing networks but more metrics, targeted at specific situations and questions, are needed.

Despite potential improvements in this direction, to gain *real* evaluative and discriminative power over network models, robust statistical methods such as maximum likelihood estimation need to be relied on. We have made a step in this direction which, hopefully, will contribute to consolidating and further developing the field of network models. Open issues include estimates of parameter uncertainty in the case of real data by bootstrap methods [32,33], identification of situations where maximum likelihood estimates converge to true parameter values (including model misspecification as in Ref. [21] which is of particular importance to parameter estimates in complex systems), and improvements of the mean-field likelihood

estimation which was introduced in Sec. III B. It needs to be stressed that the potential impact of parameter estimation far exceeds the academic problem of model validation: Model parameters, once known, can be directly useful in practice. In the case of preferential attachment with relevance decay, for example, the overall rate of relevance or interest decay is closely connected to the most successful strategy in the competition for attention [28]. On the other hand, the initial, current, or total relevance values of individual items can be used to detect which items deserve to be examined more closely.

### ACKNOWLEDGMENTS

This work was supported by the EU FET-Open Grant No. 231200 (project QLectives) and by the Swiss National Science Foundation Grant No. 200020-143272.

### APPENDIX : ON THE SHAPE OF THE LIKELIHOOD FUNCTION

Greedy sequential optimization is possible because the likelihood function in our case does not have a large number

of disparate local minima. We explain this fact for the PA-RD model which is parametrized by the initial node relevances  $I_i$  and global parameters  $A$  and  $\Theta$ . Reference [16] shows that the expected final degree of node  $i$  grows with  $I_i$  which implies that  $\mathcal{L}(\mathcal{D}|\mathbf{I}, A, \Theta)$  has a unique maximum in  $I_i$  when all other parameters are fixed. Likelihood of the data thus has a unique maximum in the space of all initial relevance values. Similar behavior can be observed for  $A$ . When  $A \rightarrow \infty$ , likelihood of the artificial data is small because the model simplifies to random attachment which is obviously at odds with the data. As  $A$  decreases, the likelihood grows but it eventually saturates and decreases when  $A$  becomes so small that new nodes cannot attract their first links. The case is different for  $\Theta$ . Its extremely small values can be easily refuted by the data as they would imply links always arriving at the latest node. On the other hand, large  $\Theta$  can be accommodated by an appropriate choice of the initial relevance values which is demonstrated by Fig. 2. To prevent the sequential updating of parameters from converging to a wrong solution, one can for example add a suitable penalization term as we do in Eq. (5).

- 
- [1] S. N. Dorogovtsev and J. F. F. Mendes, *Adv. Phys.* **51**, 1079 (2002).
  - [2] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang, *Phys. Rep.* **424**, 175 (2006).
  - [3] M. E. J. Newman, *Networks: An Introduction* (Oxford University Press, New York, 2010).
  - [4] M. C. González, A.-L. Barabási, *Nat. Phys.* **3**, 224 (2007).
  - [5] P. D. Hoff, A. E. Raftery, and M. S. Handcock, *J. Am. Stat. Assoc.* **97**, 1090 (2002).
  - [6] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. U. Boas, *Adv. Phys.* **56**, 167 (2007).
  - [7] E. D. Kolaczyk, *Statistical Analysis of Network Data* (Springer, New York, 2009).
  - [8] F. Papadopoulos, M. Kitsak, M. Á. Serrano, M. Boguñá, D. Krioukov, *Nature (London)* **489**, 537 (2012).
  - [9] M. Li, H. Zou, S. Guan, X. Gong, K. Li, Z. Di, and C.-H. Lai, *Sci. Rep.* **3**, 2512 (2013).
  - [10] M. P. H. Stumpf and M. A. Porter, *Science* **335**, 665 (2012).
  - [11] A. van den Bos, *Parameter Estimation for Scientists and Engineers* (Wiley-Interscience, Hoboken, NJ, 2007).
  - [12] D. A. Freedman, *Statistical Models: Theory and Practice*, 2nd ed. (Cambridge University Press, New York, 2009).
  - [13] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data* (Springer, Berlin, Heidelberg, 2011).
  - [14] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, In *KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, 2008), p. 462.
  - [15] J. Leskovec, D. Chakrabarti, J. Kleinberg, C. Faloutsos, and Z. Ghahramani, *J. Mach. Learn. Res.* **11**, 985 (2010).
  - [16] M. Medo, G. Cimini, and S. Gualdi, *Phys. Rev. Lett.* **107**, 238701 (2011).
  - [17] R. Albert and A.-L. Barabási, *Rev. Mod. Phys.* **74**, 47 (2002).
  - [18] Y.-H. Eom and S. Fortunato, *PLoS ONE* **6**, e24926 (2011).
  - [19] P. Holme and J. Saramäki, *Phys. Rep.* **519**, 97 (2012).
  - [20] D. Wang, C. Song, and A.-L. Barabási, *Science* **342**, 127 (2013).
  - [21] H. Owhadi, C. Scovel, and T. Sullivan, [arXiv:1308.6306](https://arxiv.org/abs/1308.6306).
  - [22] M. S. Shang, L. Lü, Y.-C. Zhang, and T. Zhou, *Europhys. Lett.* **90**, 48006 (2010).
  - [23] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. (Morgan Kaufmann, Waltham, MA, 2011).
  - [24] R. Tibshirani, *J. Roy. Stat. Soc. B* **58**, 267 (1996).
  - [25] [www.unifr.ch/econophysics](http://www.unifr.ch/econophysics).
  - [26] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevE.89.032801> for data download.
  - [27] G. Bianconi and A.-L. Barabási, *Europhys. Lett.* **54**, 436 (2001).
  - [28] B. A. Huberman, *J. Stat. Phys.* **151**, 329 (2013).
  - [29] F. Radicchi, S. Fortunato, and C. Castellano, *Proc. Natl. Acad. Sci. USA* **105**, 17268 (2008).
  - [30] K. P. Burnham and D. R. Anderson, *Model Selection and Multi-model Inference: A Practical Information-Theoretic Approach*, 2nd ed. (Springer-Verlag, New York, 2002).
  - [31] G. Claeksens and N. L. Hjort, *Model Selection and Model Averaging* (Cambridge University Press, New York, 2008).
  - [32] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application* (Cambridge University Press, Cambridge, 1997).
  - [33] C. Shalizi, *Am. Sci.* **98**, 186 (2010).